Proc. London Communications Symposium, pp 231-234, September 8th-9th, 2005.

A VISUAL ATTENTION BASED METHOD FOR OBJECT TRACKING

Shijie Zhang and Fred Stentiford
Department of Electronic and Electrical Engineering, University College London, Adastral Park Campus

Key words to describe this work: Visual attention, video surveillance, region of interest, saliency map, object tracking, motion detection, MPEG video encoder

Abstract

Object tracking is a very important operation in many surveillance applications. It is also closely related to motion detection/estimation and object recognition. This paper proposes a visual attention based method for object tracking. This technique is used to generate motion vectors for each frame in a moving video sequence. Results are compared with MPEG video encoder produced motion vectors. Further discussion of this method will be given on the extension of algorithm for 2D static images to sequences of frames in the time dimension in future work. Preliminary results are described.

1. Introduction

It is widely believed that object tracking has become a major part in the video surveillance systems. Once the object has entered the viewing field of the camera, it will attract our attention to focus on its movement before it disappears entirely from the area. Throughout the tracking operation, motion plays an important role. Motion catches our attention, and attention triggers tracking. It is well appreciated that visual attention is a vital part of our human visual system. Models of attention have been applied to static images in applications such as target location, image retrieval, auto-focus in cameras, and image compression. Current research is now active in video processing and this study will investigate attention based algorithms for object tracking in real-time scenarios with application to high performance CCTV systems for video surveillance.

Section 2 gives a brief literature survey on both visual attention algorithms and object tracking. In Section 3, our approach is presented. Results are shown in Section 4 along with some discussion. Finally, Section 5 outlines conclusions and future work.

2. Background

2.1 Visual attention algorithms

Attention algorithms identify salient regions as foreground allowing unimportant background regions to be largely ignored without significantly affecting the overall information content and perceptual quality of the image. By doing this, it can enable processing to concentrate on regions of interest by analogy with our human visual system.

The basis of many visual attention models over the last two decades is the feature integration theory of Treisman[1] that was derived from visual search experiments. Based on this theory, Koch and Ullman[2] have suggested a model that leads to the generation of a saliency map that encodes the saliency of image regions. Meaningful objects are then identified at a second stage which requires focused attention. Itti and Koch[3] have proposed a visual attention system based on the behavior and the neuronal architecture of the early primate visual system. The neurobiological model of visual attention is capable of good performance with complex natural scenes. The major strength of the approach lies in the massively parallel implementation which has low computational requirements. However the system can only work for object features explicitly represented therefore it will fail at detecting targets salient for unimplemented feature types. Also the model does not include a magnocellular motion channel which is very important in human saliency detection as it is known that motion plays an important role in visual attention. Corchs and Ciocca[4] proposed an approach to select the key frames for video summarization. The frames are selected based on the results of the analysis of the events in terms of regions of interest which are obtained from a biologically based model of visual attention. The model is the bottom-up component given by the V1 region of the primary visual cortex where regions of interest of the image are determined from the maps of neural activities of the V1 neurons. The method is effective in video summarization, but the model only works for grey-scales so that no colour information is taken into account. Avrithis[5] presented an extension of visual attention schemes in video sequences by incorporating temporal dimensions. It is expected to be used for revealing interesting events across the sequence such as occlusions and short occurrences of objects thereby providing a basis for video surveillance. The proposed framework is an extension of Itti's model[3] to the spatiotemporal space. Both intensity and colour information are used in the model which treats the video sequence as a video volume with frame number(time) being the third dimension. However, previous research work on saliency maps based on

visual attention has always used pre-selected features, which is avoided in Stentiford's model[6]. In the paper, the author proposed a novel model of visual attention applied to the automatic assessment of the degree of DNA damage in cultured human lung fibroblasts. The visual attention estimator measures the dissimilarity between neighbourhoods in the image giving higher visual attention score/value to neighbouring pixel configurations that do not match identical positional arrangements in other randomly selected neighborhoods in the image. The similarity measure approach is then later used in application for content based image retrieval[7] and image compression[8]. Wolfe[9] has introduced attention mechanisms into the visual search task which then brought in the idea of guided search. In his model, stimuli are divided into two pre-attentive processes, which are then combined into an attention-guiding activation map. However including Wolfe himself all of us are aware that two factors are ignored in this approach, one is the eye movement("overt") and attention moment("covert"), the other is the concentration of fovea on centres of regions. In short, feature attributes guide attention. It seems that features are the dominant factors in visual attention mechanisms with attention always depending on features. It is within our interest not to rely on pre-selected features using our own visual attention algorithm.

2.2 Object Tracking

In general, the existing approaches dealing with object tracking can be classified into several categories, i.e. feature-based approach, template-based methods, gradient-based methods, statistical model and prediction approach. A few present tracking methods are presented and discussed below.

In [10] a novel online feature learning approach was proposed for faster and effective object tracking. However, a pre-selected feature is required which means a priori characteristics need to be known. Also, graph-based object tracking[11] has been applied to each image in a sequence and represented in a region adjacency map, so that the object tracking becomes a graph-matching problem. The approach keeps track of occluded objects. Using hidden Markov models [12] can be a faster and low computational solution, however it is not suitable in the presence of similar objects and small/deformable objects. In [13], a linear prediction method was proposed to predict the centroid of the moving object. The high accuracy performance is superior to that of the Kalman filter, but its use is limited to single object movement.

We therefore believe a much faster, effective, adaptive and robust method is required.

3. Proposed motion tracking approach

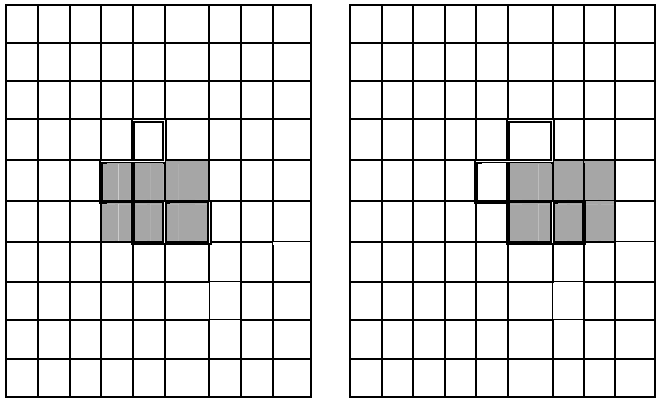


Figure 1. Two video frames (1 and 2)

In Figure 1 the grey object is moving from left to right between Frames 1 and 2. A 4 pixel fork from frame 1 is shown mismatching frame 2 after a displacement of (1,0).

The tracking algorithm is as follows:

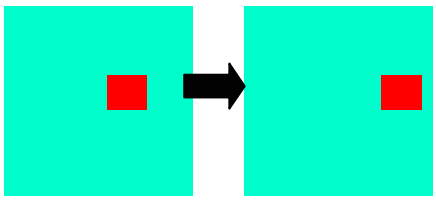
1. For a frame of dimension I x J, initialise arrays OFFSETX(I,J), OFFSETY(I,J), COUNT(I,J)
2. Generate a random n pixel fork in frame 1 (n = 4 in Figure 1) in which half the pixels mismatch the other half. In the case of Figure 1 two pixels will be inside the object and two outside
3. Apply the fork to a random position in Frame 2 with offsets dx and dy
4. If the fork matches then for each fork pixel position (i,j) in frame 1:
   OFFSETX(i,j)= OFFSETX(i,j) + dx
   OFFSETY(i,j) = OFFSETY(i,j) + dy
   COUNT(i,j) = COUNT(i,j) + 1
   Otherwise loop to step 3 N times
5. Loop to 2 M times
6. We now have two arrays for total offsets x and y. Dividing them by the counter matrix respectively, we obtain the average offset arrays for x and y.

Now we have the offset arrays we can plot a visualisation of the object motion vectors.

4. Results and discussion

4.1 Artificial object movement

The algorithm was implemented in Matlab 7 and tested on a video sequence consisting of 33x33 frames. The maximum fork size was set to 3 x 3 with N=M=1000. Two frames are shown in Figure 1 of a red square moving from left to right on the blue background.



frame 1 frame 2

Figure 2. Artificial Frames

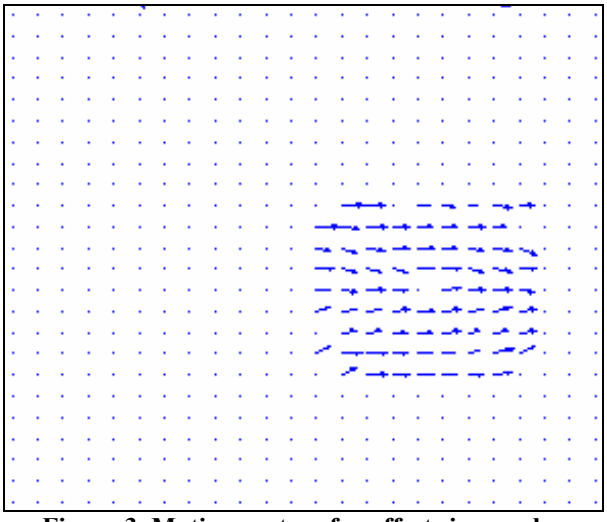
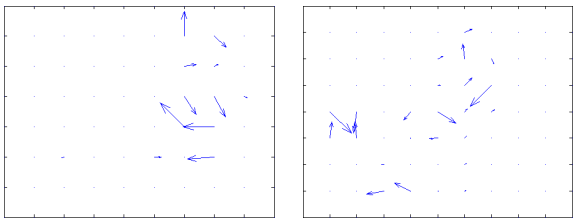


Figure 3. Motion vectors for offsets in x and y

Figure 3 shows the motion vectors derived from the motion implied by the movement between frames 1 and 2 in Figure 2 using this method. The length of the arrows is proportionate to the magnitude. The processing took approximately half a minute on a 1.7 GHz Pentium M Processor with 480 MB of RAM. Some results of the MPEG encoder generated motion vectors are shown in Figure 3[14].

I frame in MPEG video

P frame 1 **P frame 2**

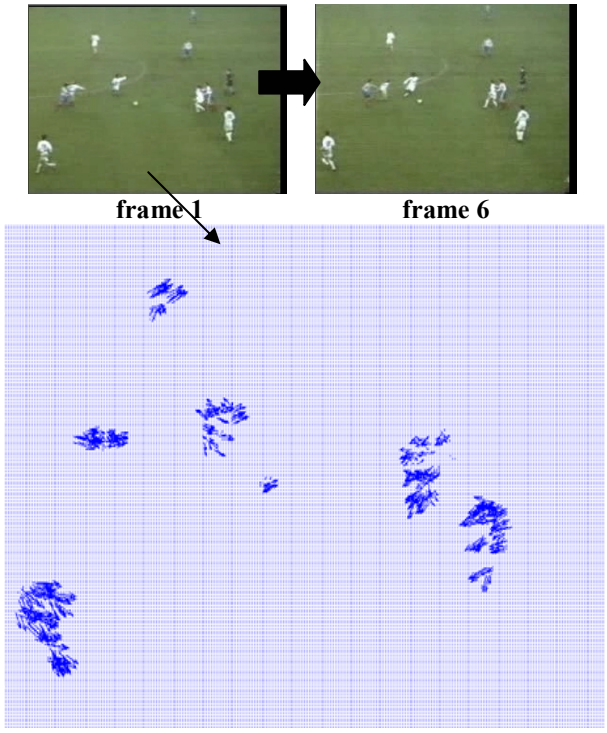


motion vectors for P 1 **motion vectors for P 2**

Figure 4. MPEG motion vectors

Although the motion vectors have been plotted out for both P frames it is hard to predict the general directions of object motion. The video using the MPEG encoder was sub-sampled so that motion vectors are only generated for P frames at a rate of 10 frames per second. Furthermore, the erratic motion vectors have to be smoothed out by filtering at a given point over time.

4.2 Real data



frame 1 **frame 6**

Figure 5. Two frames and corresponding motion vectors

Frames from a football video are shown in Figure 5 together with the extracted motion vectors. The frames are 272x400, uncompressed avi format with a frame rate of 24 fps. Not only are the motion vectors for the players on the football field plotted, but also the motion of the ball movement was also spotted in the middle of the pitch. The vector diagram has relatively low noise and low randomness. Also, some of the motion vectors still need to be looked into to distinguish different

player movements. In total, movements of 8 players and the ball have been identified by the motion vectors.

5. Conclusions

A visual attention mechanism has been proposed to be used for object tracking. The new method can extract motion vectors and initial results are promising when compared with an existing MPEG video encoder. The method extracts the object displacement between frames and may be used to compute the absolute velocity given the geometry of the scene. In future work the algorithm will be tested and refined on further data taken from a surveillance scenario.

Acknowledgement

The project is sponsored by European Commission Framework Programme 6 Network of Excellence MUSCLE(Multimedia Understanding through Semantics, Computation and Learning).
The author would also like to thank colleagues in the British Telecom Broadband Applications Research Centre.

References

[1] A.M. Treisman and G. Gelade, "A feature-integration theory of attention", Cognitive Psychology, vol. 12, no. 1, pp. 97-136, Jan. 1980.
[2] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry", Human Neurobiology, vol. 4, pp. 219-227, 1985.
[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans. on Pattern Analysis and Machine Intelligence(PAMI), vol. 20, no. 11, pp. 1254-1259, 1998.
[4] S. Corchs, G. Ciocca, R. Schettini and G. Deco, "Video summarization using a neurodynamical model of visual attention", IEEE, 2004.
[5] K. Rapantzikos, N. Tsapatsoulis and Y. Avrithis, "Spatiotemporal visual attention architecture for video analysis", IEEE, 2004.
[6] F. W. M. Stentiford, N. Morley, and A. Curnow, "Automatic identification of regions of interest with application to the quantification of DNA damage in cells," in Human Vision and Electronic Imaging VII, B. E. Rogowitz, T. N. Pappas, Editors, Proc SPIE vol. 4662, pp 244-253, San Jose, 20-26 Jan, 2002.
[7] A.Bamidele, F.W.M Stentiford and J. Morphett, "An Attention Based Approach to Content Based Image Retrieval**",** British Telecommunication Technology Journal, Intelligent Spaces- an application of pervasive ICT, vol. 22, no. 3, July 2004.
[8] F. W. M. Stentiford, "An estimator for visual attention through competitive novelty with application to image compression", Picture Coding Symposium, pp. 101-104, Seoul, 24-27 April, 2001.
[9] J. M. Wolfe, "Guided search 2.0: A revised model of visual search", Psychonomic Bulletin & Review, vol. 1, pp. 202-238, 1994.
[10] F. Liu and J.B. Su, "Reinforcement learning-based feature learning for object tracking", ICPR 2004, vol. 2, pp. 748-751, 23-26 August.
[11] C. Gomila and Fernand Meyer, "Graph-based object tracking", ICIP 2003, vol. 3, pp. II-41-4, 14-17 Sept.
[12] S. Lefevre, E Bouton, T. Brouard and N. Vincent, "A new way to use hidden markov models for object tracking in video sequences", ICIP 2003, vol. 2, pp. III-117-20, 14-17 Sept.
[13] P.Y. Yeoh and S.A.R. Abu-Bakar, "Accurate real-time object tracking with linear prediction method", ICIP 2003, vol. 2, pp. III-941-4, 14-17 Sept.
[14] M. Davenport, C. Flesher, M. Poonawala, D. Suksumrit, University of Rice, Texas, US, Group Project 2002, http://www.owlnet.rice.edu/~elec301/Projects02/motionVector/trackingresults.html.